Exploiting network analysis to create a novel sentinel surveillance system for efficient, rapid detection of emerging Clostridioides difficile strains in England

Code availability

All code will be moved to a publically accessible location once our organisation's current GitHub restructuring is complete, so that a permanent link can be supplied (replacing the placeholder in the manuscript statement).

R Scripts

The following R scripts are provided:
`1_Network_Identification.R`
`2_Spread_Simulations.R`
`3_Sentinel_Selection.R`
`4_Sentinel_Set_Evaluation.R`
`4appendix_Sentinel_Set_Evaluation.R`
These correspond to the four steps in Figure 1 of the manuscript. Each uses output from the lower numbered scripts.

Also provided is `0_Dummy_Data.R`.
This generates dummy data files. NB the facilities and network connections output by `HospitalNetwork::stay_linelist` are somewhat homogeneous (severely limiting the ability to optimise sentinel choice) so we recommend using primarily to assist with code familiarisation.

The following R packages are called within these scripts: `tidyverse`, `readr`, `janitor`, `igraph`, `tictoc`.
The manuscript analysis used R version 4.4.1, on RStudio (version 2024.04.2), with `tidyverse 2.0.0`, `readr 2.1.5`, `janitor 2.2.0`, `igraph 2.0.3`, `tictoc 1.2.1`. The manuscript analysis used a HospiNet object (see below) generated using `HospitalNetwork 0.9.3`.

Data input

The following four data input types are required (read-in by `1_Network_Identification.R`):
* HospiNet object: output from `HospitalNetwork` package (data for which is a line list of hospital spells), contains the formatted network adjacency matrix
* facility information : csv file of real world information for facilties (name, type, region)
* annual cases : csv file with annual cases by facility, used as the all-cases denominator in sampling protocols
* fitted normal : csv with normal distribution parameter values, used in estimation of specimens supplied for sequencing
NB the facility information and annual cases files include a variable `fID`, these values (alphanumeric codes for each facility) **must** match the `fID` used in the HospiNet object for each real world facility.

Analysis parameters and file-naming convention

A file-naming convention is used to capture the analysis parameter values used at each step, including the parameters tacitly inherited from an earlier stage. This enables correct identification of correct precursors and the faithful sharing of variables within an analysis pathway, such as the `fID` above. It is also intended to facilitate output file selection for the comparative aspects of sentinel set evaluation.

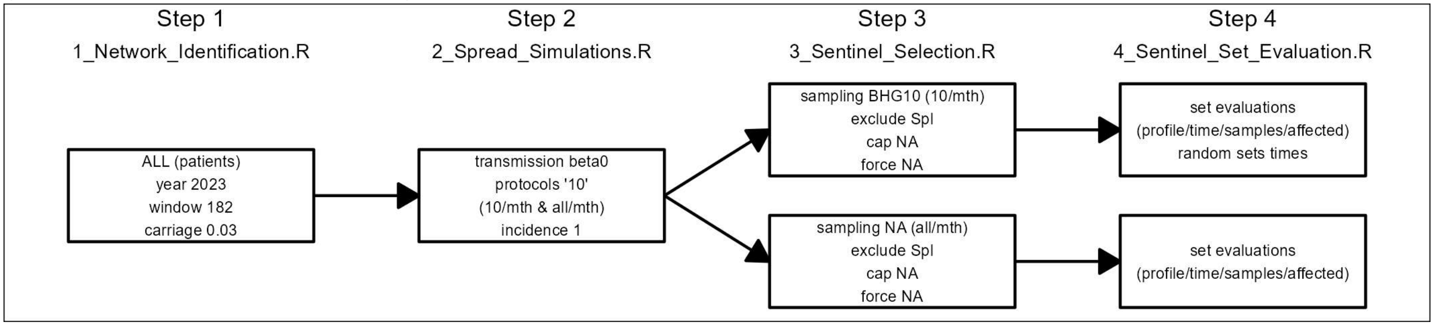
The parameters are included as a suffix within the filename, with parameters from each step separated by `_` and within each step by `-`
eg `sentinelcsv_ALL-fy2023-wdw182-cw03_p0005-sp10-i100_meanBHG10-excSpl-noCap-noFrc`

From	Suffix indicator	Parameter description
Step 1	(none)	shorthand for patients included in network data
Step 1	`fy`	financial year of network data
Step 1	`wdw`	inter-spell window (from network definition in `HospitalNetwork`)
Step 1	`cw`	carriage weighting
Step 2	`p`	transmission parameter
Step 2	`sp`	sampling protocols
Step 2	`i`	incidence multiplier
Step 3	`mean`	observations used for optimisation
Step 3	`exc`	facilities excluded from the priority list
Step 3	`cap`	facilities with capped use in the priority list
Step 3	`frc`	facilities forced to occupy top positions in the priority list

Other user-definable values
Step 1: `draws_from_distribution` number of random draws for sample burden estimation
Step 2: `max_timestep` end incomplete simulation at this timestep
Step 3: `sims_per_seed` number of simulations for each index facility
Step 4: `nRandom` number of random sets to generate
Step 4: `nMember` can be used nominate a sentinel set size to VIEW specific details (saved files include values for all set sizes)
Parameter and user-defined values are situated at the head of each script for ease of entry.

Workstream

As an example, to generate the Benchmark sentinel set in the manuscript:
* Import data (including network), apply carriage weighting (`1_Network_Identification.R`)
* Calibrate the transmission parameter `p` using evidence from simulated spread (i.e. facility affected time `atimes`), then procede using `p` = beta0 (`2_Spread_Simulations.R`)
* Generate priority lists, a list for each candidate sampling protocol included in previous step (`3_Sentinel_Selection.R`)
* Obtain evaluation data for each priority list, compare and choose sampling protocol and sentinel set size (`4_Sentinel_Set_Evaluation.R`)



Using the included ‘SMALLDUMMY’ files (taken from `0_Dummy_Data.R` output) with beta0 = 0.00001, this flowchart completes in c.30minutes. Network size and transmission parameter strongly affect completion time.

To generate optimised sets for other parameter combinations:
* branch from this pathway, start at the Step where the (first) different parameter value is introduced e.g. for a different transmission parameter `p` run Steps 2-3-4, for different window `wdw` run Steps 1-2-3-4 (reading-in the different HospiNet object).

To generate the performance of a set/list in parameter context other than where it was optimised:
* complete Steps 1-2 for the alternative context
* use `4appendix_Sentinel_Set_Evaluation.R` to save the set/list with a shorthand name, noting the `fID` codes for real world facilities must be those used in the alternative context, e.g. if time-shifting to before a merger then replace current facility with all predecessors
* “force” that set/list in selection criteria (Step 3) and complete Steps 3-4 for the alternative context.